DOCUMENT RESUME

ED 300 451 TM 012 500

AUTHOR	Green, Donald Ross
TITLE	Interpreting Scores from Standardized Achievement Tests.
PUB DATE	86
NOTE	32p.; A shorter version of this paper was included in the February 1987 "National Association of Secondary Schools Bulletin" (Volume 71, Number 496, pages 23-35).
PUB TYPE	Reports - Research/Technical (143)
EDRS PRICE	MF01/PC02 Plus Postage.
DESCRIPTORS	*Achievement Tests; Elementary Secondary Education; Equated Scores; Grade Equivalent Scores; Raw Scores; *Scoring; *Standardized Tests; *Test Interpretation; Test Use
IDENTIFIERS	Normal Curve Equivalent Scores; Percentile Ranks

ABSTRACT

Uses of the variety of scores generated by standardized achievement tests are discussed. Desirable characteristics of scales, raw score scales, percent of correct items, percentile ranks, grade equivalents, normal curve equivalents, and scale scores are considered. The various meanings and purposes of each type of score are discussed. It is concluded that, for direct and easy interpretations concerning individual and group status, national percentile ranks are probably the most effective. However, normal curve equivalents are just about as good for these purposes for those who are comfortable with standard scores. Scale scores are most effective for tracking individual and group growth patterns and for most research purposes. Percent correct scales are most useful for assessing the amount of knowledge or degree of mastery, but these scores require that the interpreter be thoroughly familiar with the content of the domain and the content of the test. Number correct scores and grade equivalent scores are the least useful; however, as long as they are not used in isolation from the others, they add information and, therefore, should be included in score reports. Sample student record sheets for grade 11 associated with the California Achievement Tests are appended. (TJH)

****
* Reproductions supplied by EDRS are the best that can be made *
* from the original document. *
****

ED300451

Interpreting Scores from Standardized Achievement Tests

U.S. DEPARTMENT OF EDUCATION
Office of Educational Research and Improvement
EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)
☑ This document has been reproduced as received from the person or organization originating it.
☐ Minor changes have been made to improve reproduction quality.
• Points of view or opinions stated in this document do not necessarily represent official OERI position or policy.

by

"PERMISSION TO REPRODUCE THIS MATERIAL HAS BEEN GRANTED BY
Donald Ross Green
TO THE EDUCATIONAL RESOURCES INFORMATION CENTER (ERIC)."

Donald Ross Green
CTB/McGraw-Hill
Monterey, California

1986*

*A shorter version of this paper appeared in the **National Association of Secondary Schools Bulletin** Volume 71, Number 496, pp. 23-35, February, 1987.

TM012500

A newly appointed principal has received the reports of scores on a battery of standardized achievement tests just administered to the students in the school. Figure 1 shows the results for some of the students in one of the 11th grade classes and Figure 2 gives a summary of the scores for that class. It is immediately obvious that several kinds of scores are reported.

Insert Figures 1 and 2 about here

To be seen in these reports are:

National percentile ranks (NP)

Grade equivalents (GE)

Normal curve equivalents (NCE)

Scale scores (SS)

Raw scores or number correct (NCR)

An expected or anticipated scale score (AASS) based on the student's performance on the Test of Cognitive Skills is also shown. Various scales other than these may appear in such reports but this list includes the most common ones.

A second kind of report for the same class indicates degree of mastery based on percent of items correct for the objectives measured in each of the content areas (Figure 3). The summary of these data for both the 11th grade classes in the school appear in Figure 4; this

---

Insert Figures 3 and 4 about here

---

report shows the percent of the students who met the criterion of mastery for each objective and an average for the class of a modified percent of items correct score.

It would be nice, but entirely unreasonable, to believe that the meaning of all the numbers displayed in these four figures will be self evident to both the principal and all the members of the staff who might use them. This article is based on the presumably more realistic assumption that explanations would be useful.

The questions considered include:

Is there any reason to report more than one kind of score?

Do they have different meanings?

Are some better than others?

Do they have different uses?

Clearly the answers to these questions are all "yes" since writers don't ask rhetorical questions that lead nowhere. Thus after making a few assumptions to avoid the many critical issues around tests that are not directly relevant to the topic of this article, some desirable characteristics of score scales will be described briefly; this will include a discussion of score distributions. Then each of the scores listed will be described and contrasted with the others, indicating their differences, uses, advantages and problems. The data in the reports will be used to illustrate some of these points.

Assumptions

To keep matters simple it will help to assume that the tests are reasonably valid for their intended uses in this school and that their reliabilities are adequate. Given tests from any one of the major published achievement batteries in current use in this country such as the one illustrated in the figures, the second assumption is certainly

safe while the first is at least possible depending on how the data are to be used.

Of course, validity questions are always critical for proper test use and they are discussed in several of the articles in this issue of the **Bulletin**, but they are not considered here. When considering the matters discussed in this article a willingness to assume the best is helpful (note the initial assertion about the timeliness of the reports). The critics of tests might note that skepticism about test scores can be better directed when it is grounded in understanding of the scales being used.

Desirable Characteristics of Scales

Although teachers tend to believe that standardized tests are given to harass them, in fact they are given to obtain information about student knowledge and skills, individually and collectively. To that end, ways are needed to enable one to tell whether scores are good or poor, whether they differ from that on another test, whether they represent an increase in knowledge since the last test for that group and/or for the comparable group the previous year, and so forth and so on.

To provide the various sorts of information needed to do these things, score scales must have some of the following qualities. It is not possible to have them all. Each of the scales being discussed has at least one of these desirable characteristics but none has more than three of them.

1. The scores should provide a rational ordering of the students with respect to the knowledge or skill being measured. Within the limits of the effectiveness of any test in discriminating (differentiating) among students with respect to these quali-

ties, all the scales being considered are adequate on this count.

2. A score should provide direct information, about the relative status of the individuals obtaining that score, i.e., their rank or position in an appropriate reference group. It should not be necessary to look at frequency distributions and the statistics to decide what the score means. Some derived scores, notably NPs but also standard scores such as NCEs are designed to do that.
3. A score should provide some direct information about what a person with that score knows or can do. In a limited sense GEs do that, i.e., they suggest a comparison group directly rather than rank in a prespecified group. Percent correct scales also do this if the tests meet certain criteria. By and large this is a desideratum of score scales that is not currently realized adequately.
4. The scale should consist of units such that statistics such as means and standard deviations can be calculated without distortions, and also such that differences in scores can be compared at all points along the scale. The ideal is a scale whose units represent equal amounts of the trait being measured (e.g., reading ability) at all points on the scale. Since the ability is an inference drawn from the test performance, this can only be approximated. It is commonly held that scale scores meet this criterion best, that standardized scores such as NCEs do so reasonably well, while NPs and GEs are considered to do so least well.
5. The scale should be continuous over the entire range of abili-

ties measured by the various levels of the battery. The scores should display status on a scale which refers directly to amount of knowledge or skill and not just status relative to a norm group. Scale scores do this and to lesser extent so do GEs. Clearly this desideratum is in direct conflict with #2 above.

6. Finally one would like to have scales that do not lead to misinterpretations and misuses of test results. While it is common to hear that GEs fare particularly poorly on this criterion, it may be noted that all scales so far devised create this problem in one form or another.

Some people add that the scores should have a particular distribution. The one commonly espoused is the normal distribution described at length in any basic measurement or statistics text (e.g., Allen & Yen, 1979). Since instruction may be expected to have effects that produce non-normal distributions, a normal distribution is not necessarily a desirable attribute of achievement test distributions.

Figure 5 shows a normal distribution and the way the several scales being discussed relate to that distribution. While test scores essentially never fall in a normal distribution unless some transformation makes them do so, such transformations have been popular for many years. The NCR scale and GE scale shown in this figure are based on the norms for the Reading Comprehension Test, Level 19, California Achievement Tests, Form C (1977); the scale score distribution for this test was normalized and the NP and NCE score intervals are dictated by this fact.

While achievement test distributions almost always tend to look vaguely normal, they are usually skewed right or left and/or are

either flatter or more peaked than a normal distribution. However the effect on the relationships among scales of these departures from normality are not usually very large as can be seen by comparing Figure 6 with Figure 5. Figure 6 is based on the distribution of scale scores for the Reading Comprehension Test, Level 20, California Achievement Tests, Form E (1985); this distribution, in contrast to that of Form C, has not been normalized.

We can now turn to the descriptions of each of these score scales. Since the basis for obtaining raw scores is the most obvious, the descriptions of the several scales will begin with them, even though they are the worst of the lot by the criteria just listed.

Raw Score Scales

A raw score is basically the number of items that the student answers correctly, e.g., the NCR scores shown in Figure 1. If one is familiar with both the test items and the students taking the test, these numbers may provide a meaningful scale. For example, teachers often use a raw score scale for interpreting their classroom tests.

However, for most uses, raw scores from a standardized test battery are less than satisfactory. One problem is the fact that the various items do not provide equal information about the trait being measured as the uneven spacing of the NCR scales in Figures 5 and 6 demonstrates. Note that teachers often recognize this fact and therefore weight their items differentially.

In fact, just counting the number of items correct does not make the best use of the information the students' responses provide about their abilities as measured by the test. It matters which items a student gets correct as well as how many, because some items deserve more weight than others. Scoring procedures that consider the pattern of

correct responses and take the weights of those items into account can be shown to provide more reliable estimates of student ability than a simple count of items right.

Pattern scoring is the basis for all the scores shown in Figure 1 except the number correct score. For example it can be seen that Student No. 1 got 33 items right on the Reading Comprehension Test; Student No. 10 got the same number right but received slightly higher normative scores because the patterns of correct answers for these two students indicated that Student No. 10 probably is more skillful than Student No. 1.

The most important problem with raw scores is that the meaning of any raw score is far from self-evident. Is the NCR of 33 obtained by Student No.1 on the Vocabulary test a good score? Certainly a principal new to the school could not tell. Note that the NP, GE, and NCE scores all indicate directly that the performance of the student on this test was about average; these normative scores answer the question.

Student No. 1 scored 23 items right on Language Mechanics and 28 items right on Language Expression; is the latter score better than the former? The normative scores, i.e., the NPs, GEs, and NCEs all indicate the contrary. Thus taken alone the number correct provides very little information; NCRs do not provide a reasonable basis for comparing performances across tests, even if the tests have the same number of items.

Especially because our principal is new to the school, the reports would be close to useless if only the number correct scores were on the reports. However any of the other scales listed above would provide the principal with some interpretable information about the per-

formance of the students.

In short the problem with raw score scales is that unless they are enlightened by information about the performance of at least one known group on the test, their interpretation is usually ambiguous. While one can report statistics about group performance in raw scores, it is more efficient to use scales based on the performance of the known reference group.

When the group can be defined, the relationships between their performance and the test are called norms. For standardized tests the group used is usually a representative sample of some defined population such as "all 10th grade students in the United States." Local norms based on the students in the district who just took the test may also be used. For most uses standardized tests need norms; the most easily interpreted scales are those that provide direct information about the relation of a score to the performance of the norm group.

A partial exception to this dictum occurs when the test is used for domain-referenced or criterion-referenced measurement. The scale used for this purpose typically is the proportion of the items answered correctly which is discussed next, but it may be noted immediately that most users of such scales find normative information very helpful.

Percent of items correct

Obviously the meaning of raw scores is enhanced by knowing how many items there are in the test. Converting number right to percent right puts all scores on a scale that provides some direct information about student performance since one can tell immediately if the score is perfect or if it is at or below chance level. Aside from this

information however, without knowledge of test difficulty and variability, interpretation of such scores has the same problems as raw score interpretation.

For example, Student No. 12 appears to be relatively weak in Language Mechanics according to Figure 3. Does only partial knowledge (P) on all the objectives in this test indicate a poor performance? For this student perhaps - note that unlike most of the scores for this student the Anticipated Scale Score of 764 is greater than the 738 obtained - but note also that the NP of 57 is still slightly above average. The normative data puts the criterion-referenced data in perspective.

As noted above, if the test has been designed as a criterion-referenced test (or domain-referenced to be more precise) one is invited to infer that the proportion correct score represents an estimate of the amount the student knows about the objective or domain that the test purports to measure. The Objectives Performance Indices shown in Figures 3 and 4 are meant to be estimates of the proportion of the items in this larger pool the student might be expected to get correct. Note that the reports in Figures 3 and 4 only give class averages for these estimates; individual indices indicating the error bands are given in a report not shown.

It may be apparent that the concept of proportion correct fits a knowledge domain better than it does a skill. The best one can do with the latter without resorting to comparatives, i.e. without using norms, is to establish an arbitrary definition of mastery or satisfactory performance as was done in the report of individual performance on the objectives (Figure 3). Even for knowledge domains, except for a small number of well defined finite domains most of which occur in the

primary grades (e.g., the alphabet, single digit addition facts), normative information is needed to help interpret the percent correct score. Thus both the reports giving criterion referenced information also provide information about the performance of the relevant norm group on these objectives.

For example, when looking at the kinds of scores shown in Figures 3 and 4, people find it helpful to know the national average, and the averages for the school, the school system or the state. The reasons for this include the fact that variations in difficulty inevitably occur even among items intended to measure the same specific objective or skill as well as the fact that for most educational objectives the number and variety of relevant items that one can possibly write is much larger than the number one can reasonably include in a test.

The summary report (Figure 4) suggests that the students in this school are similar to the national sample at this grade, but here and there are some apparent variations from the national averages and between the two classes. How much attention should one pay to these differences? The apparently stellar performance on Affixes as indicated by the +20 percent over the national average shown in Figure 4 occurs simply because many of the individual indices are near the (arbitrary) mastery point of .75. The fact that the average index does not differ strongly from the national figure and that both are just above this cut point supports this inference. Similar cautions are in order in comparing the two classes.

Without the national figures it is hard to see how our new principal could make much sense of these data. Even with them it is hard to believe that either the principal or the staff of this school would find this percent correct and "mastery" data a satisfactory report of

test performance taken alone. Of course, they are not intended to be used alone but as a supplement to reports of the normative scores.

In sum, one or another of the derived score scales that are based on normative data are generally needed for examining and interpreting the scores for a school on a standardized achievement test. The ones discussed below are:

Percentile Ranks

Grade Equivalents

Normal Curve Equivalents

Scale Scores

Percentile Ranks

Percentiles are probably the most widely used of the derived scores. One can find the percentile rank of any score for any reference group quite easily by finding the percentage of the group scoring below the midpoint of the score interval. For standardized achievement tests the reference group of greatest interest is the sample of the national population in the relevant grade that took the test. The sample is intended to represent the population in that grade in the country at that time. Another reference group of interest to a principal is the local population at that grade at the time the test was given in the school system. Percentiles rank individuals from 1 to 99 according to their status (low to high) with respect to a specified population. A percentile rank scale meets the second criterion listed above particularly well.

Other advantages include that percentile ranks are easily understood although occasionally someone confuses them with percent correct scores. They make comparing performances on different tests easy if the reference group is the same. For example a principal can easily

compare the relative strengths of the students in a class in reading and mathematics by using median percentiles as done in Figure 2.

Percentile ranks are specific to the particular test, the particular reference group used, the time when the reference group was tested, and the time when school gives the test. Note particularly the time bound nature of this scale. NPs given on reports such as in Figure 1 are adjusted to the week of the school year the test is taken (here 02). The local percentiles are specific to that time as well.

National percentiles are additionally time bound in the sense that they tend to change from year to year; in the late `60s and early `70s high school scores declined so that the same scale scores (and raw scores) on a test yielded increasing percentile ranks when given to later groups. More recently scores have started to increase again although these gains are just beginning to reach the high school grades. Thus schools that are not improving are likely to have increasingly lower percentile ranks in the next few years (the way to hide this is to use old norms).

The major liability of NPs is that they do not approach being an equal unit scale (criterion #4). The distribution of percentile ranks is rectangular, i.e., there are equal numbers of cases in the normative group in each interval by definition. No other score exhibits this pattern. The relationship of percentile ranks to other score scales can be seen in Figures 5 and 6. Both distributions show that NP units are very unequal.

In particular it can be seen that percentile rank scores represent small intervals near the middle of the distribution and large intervals at the extremes. A consequence is that differences between scores in the middle are much smaller than they seem, while differences at

the ends are much larger than they seem; it is easier for a student to improve from the 45th to the 55th percentile than from the 90th to the 95th. Only rank order statistics (e.g., medians) are therefore appropriate to use with these scores. People who calculate means of NPs or who compare differences at different points on the scale will often get misleading results. For example, consider these Vocabulary scores from the Class Record Sheet (Figure 1):

	NP		SS	
Student 10	53		784	
		} 22		} 19
Student 11	75		803	
		} 19		} 29
Student 12	94		832	

If one looked only at NPs one would infer that the difference between students 10 and 11 was as large or larger than that between 11 and 12. However the Scale Scores contradict this conclusion, and because of their interval properties the Scale Scores are more trustworthy on this count.

While these drawbacks of percentile ranks reduce their value for research purposes and for looking at growth, it is evident that these liabilities of percentile ranks are not grave for most uses and can be avoided by merely remembering what their relationship to Scale Scores looks like.

Grade Equivalents

Grade equivalent scores are also very popular. The crusade against them by various professional groups, e.g. the International Reading Association (1980), has had no apparent effect on the demand for these scores. In fact grade equivalents have much to be said for

them - in the elementary grades. They meet all of the first three criteria to some degree - in the elementary grades. However they do not meet any of the criteria, including the first one, adequately when used at the secondary level.

A grade equivalent score represents the median score obtained by students at any given time in the course of schooling. An implicit assumption of those who use this scale is that, on the average, performance increases throughout the years of school. It is assumed that grade 2 students do better than grade 1 students and so forth. For those subjects or skills that are taught systematically and steadily through the grades such an assumption seems reasonable.

However, that assumption breaks down for most elementary school subjects and skills somewhere around the 7th or 8th grade with the consequence that growth in these basic skills then begins to slow down and the scale slowly grinds to a halt. In the upper grades the measurement of growth and change in these skills on a grade scale becomes tenuous because the overlap in content is very large and the average differences between grades are small. Figures 5 and 6 show what happens in reading comprehension.

Not only does the growth in these scores almost vanish in the high school grades but their meaning disappears as well. A grade equivalent score on a mathematics computation test makes no sense for an 11th grade student who is taking second year algebra. Note that no one advocates developing grade scores for algebra tests. Even at the lower end of the scale the apparent meaning of the score is misleading. Figure 6 shows that an 11th grade student scoring at the 5th percentile in reading comprehension has grade equivalent score of about 6.6. Does that mean that these students should be instructed with 6

grade basals? Recognizing that these students have probably been exposed to such materials for years by the time they reach the 11th grade may make it clear why this conclusion is not appropriate.

It must be concluded that grade equivalent scores have limited use and meaning for interpreting test performance at the high school level. Certainly they should not be used alone; other scores are needed to put them in perspective.

Normal Curve Equivalents

Normal curve equivalents are a form of standard score. The simplest version of a standard score scale is obtained by subtracting the arithmetic mean of the raw scores for a group from each score and dividing the result by the standard deviation. In this form the scores do indicate the status of an individual in the group albeit less clearly than NPs do, but otherwise they have all the problems of the raw scores themselves.

For many years it has been common practice to take the distribution of raw scores and transform it into a normal distribution thus changing the relationship between score points to fit this hypothesized distribution. Traditionally the mean score was set at 50 (or 100 or 500) and the standard deviation at 10 (or 15 or 100). If one normalizes the distributions of all the tests in a battery this way, the scores on different tests, and on the same test at different times, are referred to this common scale, an obvious advantage.

A few years ago it was proposed that school systems report their Title I (now Chapter I) scores using this sort of scale but with a mean of 50 and a standard deviation slightly over 21. The result is a scale in which the scores of 1, 50, and 99 have percentile ranks of 1, 50, and 99; none of the other values coincide (Talmadge & Wood, 1976).

The scores on this scale are called normal curve equivalents (NCEs) and are currently used for federal reporting purposes.

The units of a normalized scale are based on the standard deviation of the normal curve; such units are equal to the extent that the underlying trait is normally distributed. As noted earlier this not likely to be the case but it may not be far off in many instances. Figure 6 illustrates a typical case. It can be seen that the NCE scale comes closer to being spread evenly in relation to the Scale Score scale than the NP, GE, or NCR scales. Thus treating it as an equal unit scale is usually acceptable.

Obviously NCEs are the score of choice for Chapter I reporting and as noted above they have many merits since they largely satisfy criteria 1, 2, and 4. One disadvantage is that the assumption of normality may not be reasonable for one or more of the tests in the battery. Another problem is that they can easily be confused with percentiles.

Scale Scores

Scale scores are a special kind of standard score designed to provide a continuous interval scale across the several levels of a test battery. There are two procedures for developing these scales in common use today. One of them involves normalizing the distributions at each level and fitting them together, the other uses item response theory (IRT) to scale each level and to fit them together. The first makes the assumption of normality at each level which has the advantage of giving each level the same distribution. The second has the advantage that it does not make the shaky assumption of normality of the distribution and that the assumptions about the behavior of the items it does make can be checked empirically.

Scale scores more nearly meet the ideal of an equal interval scale

than the other scales available, and because they run continuously from the lowest to the highest level of the battery they make it possible to do a number of things that cannot be done with other scales. In particular it is possible to track growth over time, to compare scores from different levels of the battery and to combine sets of scores from different levels into common reports such as those illustrated. Because students who score low on, say, Level 20 of CAT/E, can be expected to get about the same scale score on Level 19, both levels of the test can be used in a class or school and the results combined into a single report; this would not be possible for the objectives performance reports since the objectives differ from level to level.

Because of the characteristics of scale scores most of the statistics included in score reports such as means and standard deviations are calculated using scale scores and then translated into NPs, GEs, and NCEs. In fact these latter scales are developed using scale score distributions. In this sense scale scores may be considered the basic scale of the test battery and the others as convenient ways of interpreting that scale.

These strong advantages are balanced by the lack of apparent meaning, internal or external, communicated directly by these numbers. Even when one becomes familiar with them, the fact the scales vary from test to test means that other scores are needed as well.

Conclusions

One may conclude that there is indeed a reason to include more than one kind of score for standardized tests. Clearly they do have different meanings and they serve different purposes. Also some are better than others but none of them satisfy all the criteria considered desirable.

For direct and easy interpretations concerning individual and group status, NPs are probably the best. However NCEs are just about as good for these purposes for those who are comfortable with standard scores. NPs tend to be the least troublesome for public reporting although nothing is trouble free for that use! None of the other scales is really satisfactory for these uses.

For tracking growth patterns of individuals and groups, and for research purposes of most sorts, scale scores are probably the best. For research purposes NCEs are often useful also. For these uses the quality of equal units is usually primary and these two scales approximate this criterion more closely than do the other scales.

For the purpose of assessing amount of knowledge or degree of mastery, percent correct scales are most useful but they require that the interpreter be thoroughly familiar with the content of the domain and the content of the test.

Number correct scores and grade equivalent scores are the least useful but, it may argued that as long as they are not used in isolation from the others, they add information and therefore should be included in score reports.

While it is rare for school systems to ask for all of these scores in their reports it is done (the examples here are real) and it is not unreasonable. In fact it is fair to conclude that more is better than less; that is, score reports that include scores from several of these scales provide school personnel with fuller information than reports limited to one kind.

Readers interested in more thorough discussions of the nature of test scales will find it useful to read Flanagan (1951), or Angoff (1971), or Chapter 7 in Allen & Yen (1979), or all of them.

References

Allen, M.J. & Yen, W.M. (1979). Introduction to measurement theory. Monterey, CA: Brooks-Cole.

Angoff, W.H. (1971). Scales, norms, and equivalent scores. In R.L. Thorndike, R.L. (Ed.), Educational Measurement (sec. ed). Washington, D.C.: American Council on Education.

CTB/McGraw-Hill (1977). California Achievement Tests, Form C. Monterey, CA.: Author.

CTB/McGraw-Hill (1985). California Achievement Tests, Form E. Monterey, CA.: Author.

Flanagan, J.C. (1951). Units, Scores, and Norms. In E.F. Lindquist (Ed.), Educational Measurement. Washington D.C.: American Council on Education.

International Reading Association (1980, June). Reading Today.p.1.

Talmadge, G.K., & Wood, C.T. (1978). User's guide: ESEA Title I evaluation and reporting system. Mountain View, CA: RMC Research Corporation.

CLASS RECORD SHEET

Figure 1

CLASS : TEACHER NO 1

SCHOOL: SCHOOL NO. TWO
DIST : DISTRICT # 1

CITY/ST: CITY NO 16 ,ST

GRADE: 11.0

TEST DATE: 09/19/85

100506

Students

CALIFORNIA ACHIEVEMENT TESTS FORMS E AND F

Students	FORM	LEVEL	SCORES	READING: READ VOCAB	READING: READ COMP	READING: READ TOTAL	LANGUAGE: LANG MECH	LANGUAGE: LANG EXPR	LANGUAGE: LANG TOTAL	MATHEMATICS: MATH COMP	MATHEMATICS: MATH C & A	MATHEMATICS: MATH TOTAL	TOTAL BATTERY †	SPELL	STDY SKILL	SCI	SOCL STDY
STUDENT NO ,1	V	E 20	NP	49	45	48	49	35	40	17	20	19	32	51	33	12	68
DATE OF BIRTH: 04/22/69			GE	11.1	10.9	11.0	10.4	8.8	9.3	7.8	7.8	7.8	8.9	11.3	8.8	6.4	12.9
CA: 16- 5			NCE	50	47	49	50	42	45	30	32	31	40	50	41	25	60
SP CODES: 012345678998765			SS	782	772	777	729	737	733	779	740	760	757	763	750	680	771
			AASS	778	765	771	714	725	720	788	752	770	756	753	748	733	729
			NCR	33	32	65	23	28	51	14	23	37	153	22	15	14	28
STUDENT NO ,10	E	E 20	NP	53	45	49	36	47	41	42	24	31	38	28	25	22	27
DATE OF BIRTH: 12/17/68			GE	11.4	10.9	11.1	9.2	10.4	9.4	9.8	8.2	8.8	9.5	9.8	8.1	7.7	8.6
CA: 16- 9			NCE	52	47	50	43	48	45	46	35	40	43	38	36	34	37
SP CODES: 012345678998765			SS	784	772	778	716	752	734	803	746	775	762	751	740	706	721
			AASS	775	766	771	723	738	731	801	764	783	762	758	757	726	726
			NCR	33	31	64	18	35	53	23	21	44	161	19	12	15	16
STUDENT NO ,11	A	E 20	NP	75	97	84	80	88	85	70	82	77	84	68	93	94	89
DATE OF BIRTH: 09/18/69			GE	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9
CA: 16- 0			NCE	64	91	71	68	74	72	61	69	66	71	60	80	82	76
SP CODES: 012345678998765			SS	803	798	801	764	815	790	823	807	815	802	772	827	796	802
			AASS	809	780	793	756	781	769	827	805	816	791	769	800	806	807
			NCR	39	49	88	27	47	74	36	41	77	239	27	23	35	33
STUDENT NO ,12	R	E 20	NP	94	89	93	57	90	79	56	69	63	80	65	86	98	77
DATE OF BIRTH: 12/08/68			GE	12.9	12.9	12.9	12.3	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9	12.9
CA: 16- 9			NCE	82	76	81	54	77	67	53	61	57	68	58	73	93	66
SP CODES: 012345678998765			SS	832	790	811	738	821	780	813	793	803	798	778	813	813	783
			AASS	817	785	803	764	797	783	825	803	815	802	777	804	775	783
			NCR	47	45	92	22	47	69	31	39	70	231	26	23	37	31
STUDENT NO ,13	M	E 20	NP	86	97	90	91	99	98	34	65	52	89	76	93	92	89
DATE OF BIRTH: 09/16/69			GE	12.9	12.9	12.9	12.9	12.9	12.9	8.9	12.9	11.5	12.9	12.9	12.9	12.9	12.9
CA: 16- 0			NCE	73	91	77	78	99	93	41	58	51	76	65	81	88	76
SP CODES: 012345678998765			SS	816	798	807	783	874	829	796	789	793	809	777	830	794	803
			AASS	809	781	796	759	784	773	825	800	813	796	772	799	790	794
			NCR	44	49	93	29	54	83	21	38	59	235	29	22	35	32
STUDENT NO ,14	F	E 20	NP	41	18	32	14	29	21	16	19	17	21	11	46	58	54
DATE OF BIRTH: 11/14/66			GE	10.2	8.7	9.4	5.0	8.1	6.5	7.7	7.7	7.7	7.9	7.2	10.4	12.9	12.9
CA: 18-10			NCE	45	31	40	27	38	33	29	31	30	33	24	48	54	52
SP CODES: 012345678998765			SS	775	758	767	686	728	707	777	738	758	744	734	765	752	755
			AASS	787	771	780	728	750	741	797	766	783	769	762	762	721	724
			NCR	29	28	57	14	27	41	16	19	35	133	13	16	25	22

TEST OF COGNITIVE SKILLS

SCORES	SEQ	ANA	MEM	VR	TOT	CSI & CONF. RANGE
NPG	10	60	15	51	24	91
NPA	11	67	19	57	29	88-95
SS	587	745	591	729	663	
						LEVEL 5
NPG	33	53	38	29	33	94
NPA	36	58	29	35	35	91-97
SS	691	728	641	685	686	
						LEVEL 5
NPG	88	*99	51	70	89	124
NPA	91	*99	64	80	93	115-134
SS	817	979	699	762	814	
						LEVEL 5
NPG	48	68	85	90	81	115
NPA	51	72	75	90	83	110-121
SS	724	762	840	835	790	
						LEVEL 5
NPG	62	77	68	88	73	115
NPA	73	86	69	91	83	111-119
SS	751	784	728	823	772	
						LEVEL 5
NPG	19	42	63	46	40	93
NPA	14	40	52	40	33	90-97
SS	642	702	740	721	701	
						LEVEL 5

QUARTER MONTH: 02 PATTERN (IRT)
CAT FORM(S): E
NP : NATIONAL PERCENTILE
NCE : NORMAL CURVE EQUIVALENT
AASS : ANTICIP. SCALE SCORE

CA : CHRONOLOGICAL AGE
CAT LEVEL(S): 20
GE : GRADE EQUIVALENT
SS : SCALE SCORE
NCR : NUMBER CORRECT

† TOTAL BATTERY INCLUDES TOTAL READING, TOTAL LANGUAGE, AND TOTAL MATHEMATICS

* MAXIMUM/MINIMUM SCORE
X NO SCORE AVAILABLE

NPG : NATL PERCENTILE BY GRADE
SS : SCALE SCORE

TCS LEVEL(S): 5
NPA : NATL PERCENTILE BY AGE
CSI : COGNITIVE SKILLS INDEX
CTBID: B111110001-03-028-000014

COPYRIGHT © 1985 BY MCGRAW HILL INC. ALL RIGHTS RESERVED

CLASS RECORD SUMMARY SHEET

CLASS : TEACHER NO 1

Figure 2

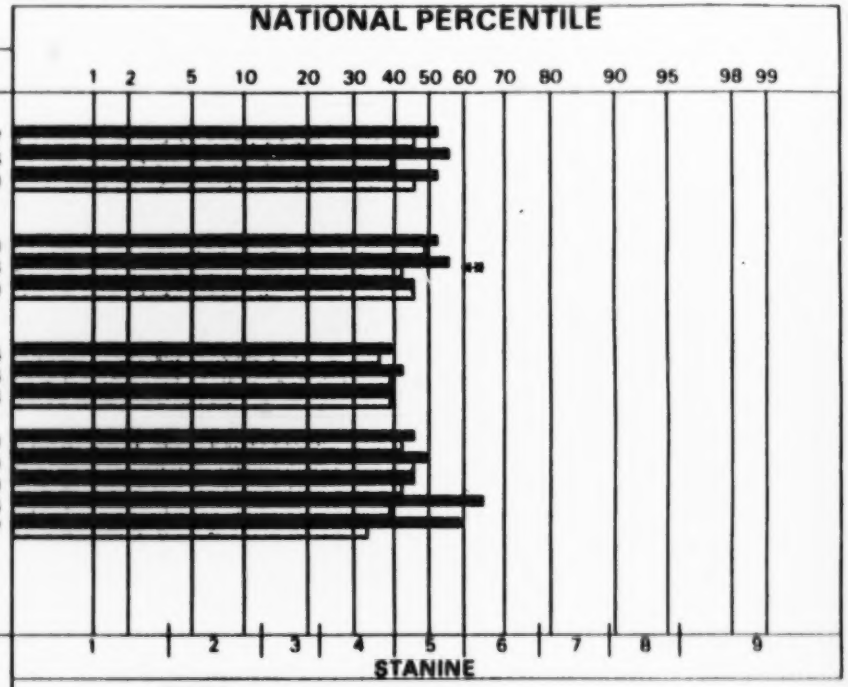
SCHOOL: SCHOOL NO. TWO
DIST : DISTRICT NO 1

CITY/ST : CITY NO 16 ST

GRADE : 11.0

TEST DATE : 09/19/85

NUMBER OF STUDENTS:		MNS	MNCE	MSS	MAASS	MDANP	MDNP
READING							
VOCABULARY	14	5.4	54.6	788.5	787.9	47.0	51.7
COMPREHENSION	14	5.5	55.5	776.2	771.8	48.0	56.5
TOTAL READING	14	5.6	54.6	782.6	780.3	46.7	52.0
LANGUAGE							
LANGUAGE MECHANICS	14	5.3	53.2	736.4	735.3	49.0	53.0
LANGUAGE EXPRESSION	14	5.8	58.4	776.1	754.2	44.0	56.5 **
TOTAL LANGUAGE	14	5.6	56.4	756.5	745.5	45.5	48.0
MATHEMATICS							
COMPUTATIONS	14	4.7	45.6	801.7	807.7	37.5	42.3
MATH CONCEPTS & APPL	14	4.9	48.7	771.6	775.9	40.0	43.5
TOTAL MATHEMATICS	14	4.7	47.5	787.0	792.1	39.0	39.5
TOTAL BATTERY	14	5.2	52.9	775.2	773.2	43.5	46.0
SPELLING	14	5.1	50.0	762.2	762.9	47.0	51.0
STUDY SKILLS	14	5.1	51.6	770.9	771.9	43.5	48.0
SCIENCE	14	5.7	57.9	753.1	752.1	39.5	64.5
SOCIAL STUDIES	14	5.4	54.6	758.2	754.6	34.0	58.7



MNS : MEAN NATIONAL STANINE
MSS : MEAN SCALE SCORE
MDANP: MD ANT. ACH. NATL PERCENTILE
MNCE : MEAN NORMAL CURVE EQUIVALENT
MAASS: MEAN ANTICIPATED ACH. SCALE SCORE
MDNP : MEDIAN NATIONAL PERCENTILE
** SIGNIFICANT DIFFERENCE

OBTAINED ACHIEVEMENT
ANTICIPATED ACHIEVEMENT

NARRATIVE SUMMARY:

THE RANGE OF OBTAINED PERCENTILE RANKS FOR THIS CLASS VARY FROM A HIGH OF 65 PERCENTILE TO A LOW OF 42 PERCENTILE. THE OBTAINED SCORE FOR THE TOTAL BATTERY INDICATES THAT THE STUDENTS ARE AVERAGE COMPARED WITH STUDENTS IN THE NATIONAL NORM GROUP.

THERE ARE STATISTICALLY SIGNIFICANT DIFFERENCES BETWEEN OBTAINED AND ANTICIPATED SCORES IN ONE TEST(S) OR TOTAL(S). THE OBTAINED SCORE IS SIGNIFICANTLY ABOVE THE ANTICIPATED SCORE FOR LANGUAGE EXPRESSION. THIS REPORT IS A SUMMARY ONLY. FOR A COMPLETE DESCRIPTION OF THE PERFORMANCE, INDIVIDUAL SCORES SHOULD BE EXAMINED.

24

TEST OF COGNITIVE SKILLS

LEVEL : 5	MDNPG	MSS
SEQUENCES	35.0	701.2
ANALOGIES	60.5	757.5
MEMORY	55.5	698.5
VERBAL REASONING	37.0	729.3
TOTAL	40.5	721.6

MEAN CSI: 102 MEAN CA: 201

MDNPG: MD NATL PERCENT BY GRD MSS : MEAN SCALE SCORE

25

QTR MTH: 02 PATTERN (IRT) FORM/LEVEL : E/20 CTBID : B111110001-03-02-000004

COPYRIGHT © 1985 BY MCGRAW HILL, INC. ALL RIGHTS RESERVED

OBJECTIVES PERFORMANCE REPORT

Figure 3

CLASS : TEACHER NO 1

SCHOOL: SCHOOL NO, TWO
DIST : DISTRICT 0 1

CITY/STATE: CITY NO 16 ST
FORM/LEVEL: E/20
GRADE: 11.0

TEST DATE: 09/19/85

100506
100 | 506

NUMBER OF STUDENTS = 14

	STUDENTS MASTERING LOCAL %	NATL %	DIFF	AVG OPI*	STUDENT NO 1 V	2 E	3 A	4 R	5 M	6 F	7 L	8 L	9 J	10 N	11 L	12 E	13 L	14 J
VOCABULARY																		
24 SYNONYMS	29	33	- 4	.67	P	P	+	+	+	P	+	P	-	P	P	P	P	P
25 ANTONYMS	0	5	- 5	.39	-	-	-	P	P	-	P	-	-	-	-	-	-	-
26 HOMONYMS	0	9	- 9	.49	-	-	P	P	P	-	P	-	-	-	-	-	P	P
27 AFFIXES	93	73	+ 20	.86	+	+	+	+	+	+	+	+	P	+	+	+	+	+
28 WORDS IN CONTEXT	36	27	+ 9	.43	P	P	+	+	+	P	+	P	-	P	P	-	+	P
COMPREHENSION																		
32 PASSAGE DETAILS	29	11	+ 18	.61	P	P	+	+	+	+	P	P	-	P	-	-	P	P
35 CHARACTER ANALYSIS	64	46	+ 18	.79	P	P	+	+	+	P	+	+	-	+	P	+	+	+
36 CENTRAL THOUGHT	14	11	+ 3	.52	-	-	+	P	+	-	P	P	-	P	-	-	-	P
37 INTERPRETING EVENTS	21	15	+ 6	.61	P	P	+	+	+	P	P	P	-	P	-	-	P	P
38 FORMS OF WRITING	64	55	+ 9	.72	P	+	+	+	+	-	+	+	P	+	P	-	+	+
39 WRITING TECHNIQUES	57	53	+ 4	.74	P	P	+	+	+	-	+	+	P	+	P	-	+	+
LANGUAGE MECHANICS																		
43 PRONOUN I,NOUN,ADJECT	36	42	- 6	.72	P	P	+	P	+	P	+	P	P	P	+	P	P	+
44 BEGINNING WORD,TITLE	29	26	+ 3	.59	P	P	+	P	+	-	+	P	-	-	+	-	P	P
46 COMMA,COLON,SEMI,QUOT	29	24	+ 5	.63	P	-	+	P	+	-	+	P	+	-	P	-	P	P
47 PROOFREADING	29	25	+ 4	.61	P	-	+	P	+	-	+	P	P	-	+	-	P	P
LANGUAGE EXPRESSION																		
49 PRONOUNS	36	29	+ 7	.73	P	P	+	+	+	P	+	P	P	P	+	P	P	P
50 VERBS	29	10	+ 19	.60	-	-	+	+	+	-	+	P	-	P	P	-	P	P
51 ADJECTIVES,ADVERBS	93	74	+ 19	.90	+	+	+	+	+	+	+	+	P	+	+	+	+	+
53 SENTENCE PATTERNS	21	8	+ 13	.45	-	-	+	P	+	-	+	-	-	-	-	-	-	-
54 SENTENCE RECOGNITION	79	66	+ 13	.86	+	+	+	+	+	P	+	+	P	+	+	P	+	+
55 SENTENCE CLARITY	29	30	- 1	.60	-	P	-	+	+	-	+	P	-	P	+	-	P	P
56 SENTENCE COMBINING	71	54	+ 17	.81	P	+	+	+	+	P	+	P	P	+	+	P	+	+
57 TOPIC SENTENCE	36	32	+ 4	.66	-	P	+	+	+	-	+	P	-	P	+	-	P	P

NAT % - NATIONAL REFERENCE GROUP 11.1
AVG OPI — AVERAGE OBJECTIVE PERFORMANCE INDEX

0-49 : NON MASTERY (-)
50-74 : PARTIAL KNOWLEDGE (P)
75-99 : MASTERY (+)
** : 100%

QTR NO: 02 PATTERN (IRT)
CTBID: B111110001-03-028-000003

COPYRIGHT © 1985 BY MCGRAW-HILL INC ALL RIGHTS RESERVED

OBJECTIVES PERFORMANCE SUMMARY

Figure 4

SCHOOL: SCHOOL NO. TWO DIST : DISTRICT NO 1

CITY/ST: CITY NO 16 ST
FORM/LEVEL: E/20 TEST DATE: 09/19/85
GRADE: 11

	PERCENT OF STUDENTS MASTERING OBJECTIVE					AVERAGE OBJECTIVES PERFORMANCE INDEX				
	TOTAL SCHOOL	NATIONAL *	DIFFERENCE	TEACHER NO 1	TEACHER NO 2	TOTAL SCHOOL	NATIONAL **	DIFFERENCE	TEACHER NO 1	TEACHER NO 2
NUMBER OF STUDENTS :	27			14	13	27			14	13
VOCABULARY										
24 SYNONYMS	33	33	0	29	38	.68	.63	+.05	.67	.70
25 ANTONYMS	0	5	- 5	0	0	.48	.38	+.02	.39	.41
26 HOMONYMS	0	9	- 9	0	0	.50	.46	+.04	.49	.51
27 AFFIXES	93	73	+ 20	91	92	.85	.80	+.05	.86	.84
28 WORDS IN CONTEXT	26	27	- 1	36	15	.64	.59	+.05	.63	.65
COMPREHENSION										
32 PASSAGE DETAILS	19	11	+ 8	29	8	.58	.54	+.04	.61	.54
35 CHARACTER ANALYSIS	56	46	+ 10	64	46	.74	.68	+.06	.79	.70
36 CENTRAL THOUGHT	15	11	+ 4	14	15	.52	.49	+.03	.52	.51
37 INTERPRETING EVENTS	19	15	+ 4	21	15	.59	.55	+.04	.61	.58
38 FORMS OF WRITING	56	55	+ 1	64	46	.71	.69	+.02	.72	.71
39 WRITING TECHNIQUES	52	53	- 1	57	46	.73	.72	+.01	.74	.72
LANGUAGE MECHANICS										
43 PRONOUN I,NOUN,ADJECT	41	42	- 1	36	46	.72	.68	+.04	.72	.71
44 BEGINNING WORD,TITLE	22	26	- 4	29	15	.60	.61	-.01	.59	.61
46 COMMA,COLON,SEMI,QUOT	19	24	- 5	29	8	.60	.58	+.02	.63	.58
47 PROOFREADING	19	25	- 6	29	8	.60	.58	+.02	.61	.59
LANGUAGE EXPRESSION										
49 PRONOUNS	30	29	+ 1	36	23	.69	.66	+.03	.73	.64
50 VERBS	19	10	+ 9	29	8	.55	.52	+.03	.60	.50
51 ADJECTIVES,ADVERBS	81	74	+ 7	93	69	.83	.81	+.02	.90	.76
53 SENTENCE PATTERNS	11	8	+ 3	21	0	.38	.38	.00	.45	.31
54 SENTENCE RECOGNITION	74	66	+ 8	79	69	.80	.77	+.03	.86	.73
55 SENTENCE CLARITY	26	30	- 4	29	23	.57	.58	-.01	.60	.54
56 SENTENCE COMBINING	63	54	+ 9	71	54	.76	.74	+.02	.81	.71
57 TOPIC SENTENCE	37	32	+ 5	36	38	.63	.59	+.04	.66	.58

* NATIONAL PERCENT MASTERY FOR OBJECTIVES 11.1
** NATIONAL AVERAGE OBJECTIVE PERFORMANCE INDEX

QUARTER MONTH: 02 IRT
CTB-ID: 0111110001-02-015

COPYRIGHT ©1985 BY MCGRAW-HILL, INC. ALL RIGHTS RESERVED.

Figure 5

Distribution of Scores on the Reading Comprehension Test,

California Achievement Tests, Form C,

Level 19, Grade 11, Fall Norms

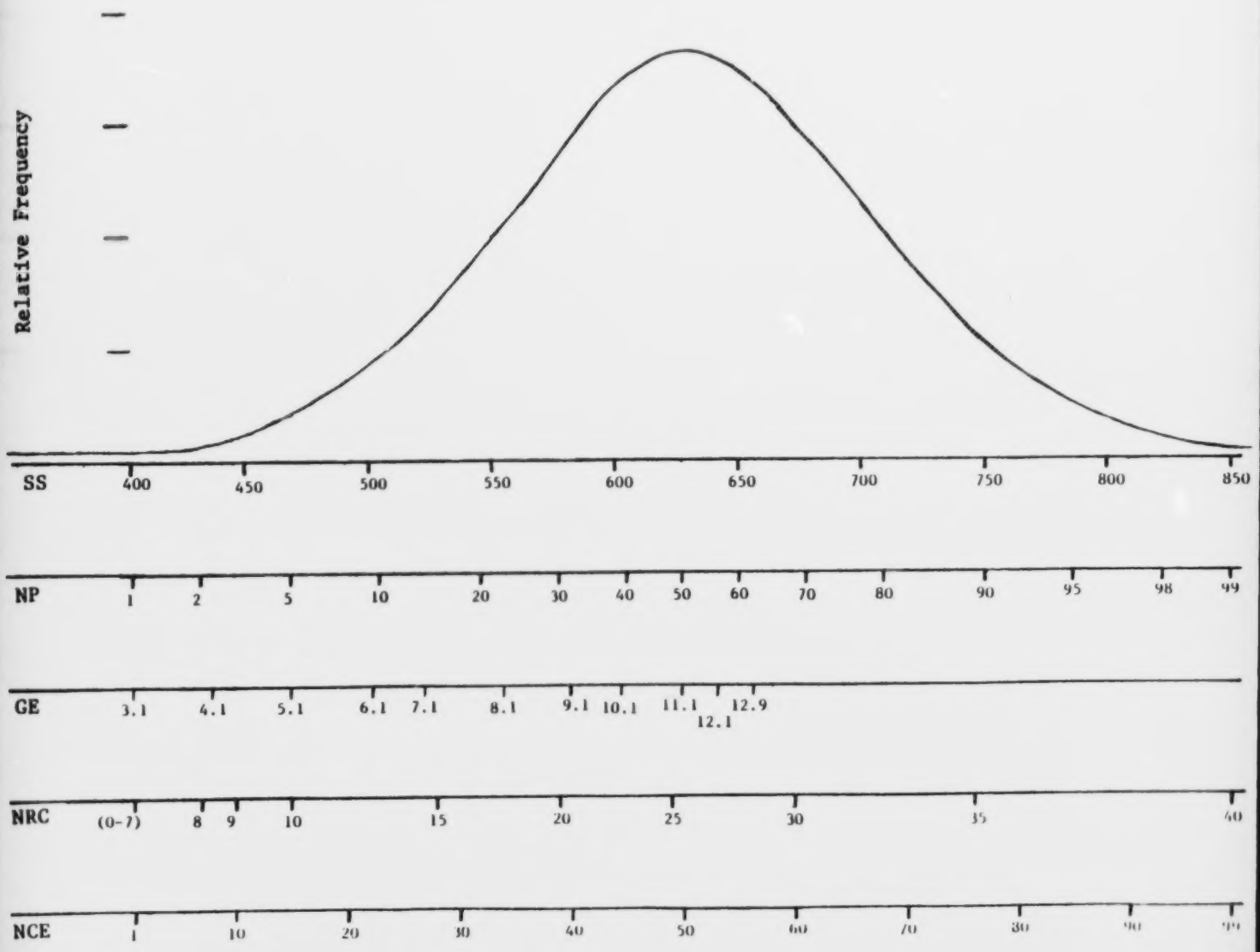


Figure 6

Distribution of Scores on the Reading Comprehension Test,

California Achievement Tests, Form E

Level 20, Grade 11, Fall Norms